***Report: Analysis and Predictive Modeming of TMDB Movie Dataset***

***Presented by Bug Smashers***

Group Members:

* Muzammil Hussain
* Ali Haider
* Jahanzaib Ejaz
* Asfand Kashif

1. Introduction

This project involves exploring the TMDB movie dataset, performing data preprocessing, analyzing trends, and building predictive models to estimate movie revenue. The dataset includes comprehensive information about movies, such as titles, genres, release dates, ratings, cast, and crew.

2. Data Exploration

The initial step in our project was to understand the structure and features of the TMDB dataset. This involved examining the dataset's various attributes, such as:

* Titles
* Genres
* Release Dates
* Ratings
* Cast and Crew

Key Functions Used:

* .merge(): Combined two Data Frames based on keys.
* .shape: Provided the dimensions of the Data Frame.
* .columns: Listed column names.
* .info(): Summarized data types, non-null values, and memory usage.

3. Data Processing and Cleaning

Data Collection and Integration

We gathered data from multiple sources and integrated them into a unified dataset, resolving inconsistencies in formats and structures.

Data Transformation

Raw data was converted into a format suitable for analysis, involving normalization and aggregation.

Data Cleaning

* Handling Missing Values: Imputation or deletion of incomplete records.
* Removing Duplicates: Ensured uniqueness of data points.

4. Exploratory Data Analysis (EDA)

Visualization techniques were used to understand patterns, trends, and relationships within the dataset.

Types of Visualizations:

* Scatter Plot: Showed relationships between two numerical variables.
* Histogram: Represented the distribution of numerical data.
* Bar Chart: Illustrated frequency, distribution, or relationship of different categories.

5. Encoding Categorical Variables

Categorical variables were converted into numerical formats for machine learning algorithms.

6. Dealing with Outliers

Outliers, which can significantly affect statistical analysis and model performance, were identified and managed appropriately.

7. Normalization and Scaling

Features were rescaled to have a mean of 0 and a standard deviation of 1, crucial for certain machine learning algorithms.

8. Feature Selection and Extraction

Feature Selection:

Selected a subset of the original features based on their importance.

Feature Extraction:

Transformed the data into a new feature space.

9. Model Building and Evaluation

Various machine learning models were built to predict movie revenue.

Models Used:

* **Decision Tree Regressor**: Built a tree structure for predictions.
* **K-Nearest Neighbors Regressor**: Predicted target values based on nearest neighbors.
* **Random Forest**: An ensemble method combining multiple decision trees.
* -**Support Vector Machine (SVM):** Effective in high-dimensional spaces, robust against overfitting.
* Model Evaluation Metrics:
* -**Mean Absolute Error (MAE**): Average of absolute differences between actual and predicted values.
* -**Mean Squared Error (MSE**): Average of the squares of the errors.
* -**Root Mean Squared Error (RMSE)**: The square root of MSE, providing an error metric in the same units as the target variable.
* **-R-squared (R²**): Proportion of variance in the dependent variable predictable from the independent variables.

10. Conclusion

The project successfully explored the TMDB dataset, handled data reprocessing, analysed trends, and built predictive models with significant accuracy. The insights gained can help in understanding movie revenue drivers and improving predictive performance.

***Thank you!***

LINK FOR CODE:

https://jupyter.org/try-jupyter/notebooks/index.html?path=py%284%29.ipynb